



JOHANNES GUTENBERG  
UNIVERSITÄT MAINZ

# Homology 3D modeling

Miguel Andrade

Faculty of Biology,

Institute of Organismic and Molecular Evolution

Johannes Gutenberg University

Mainz, Germany

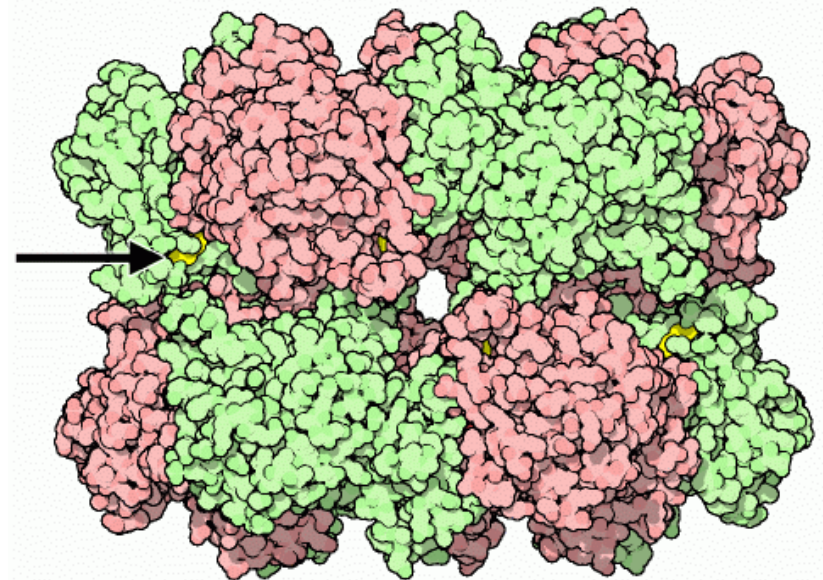
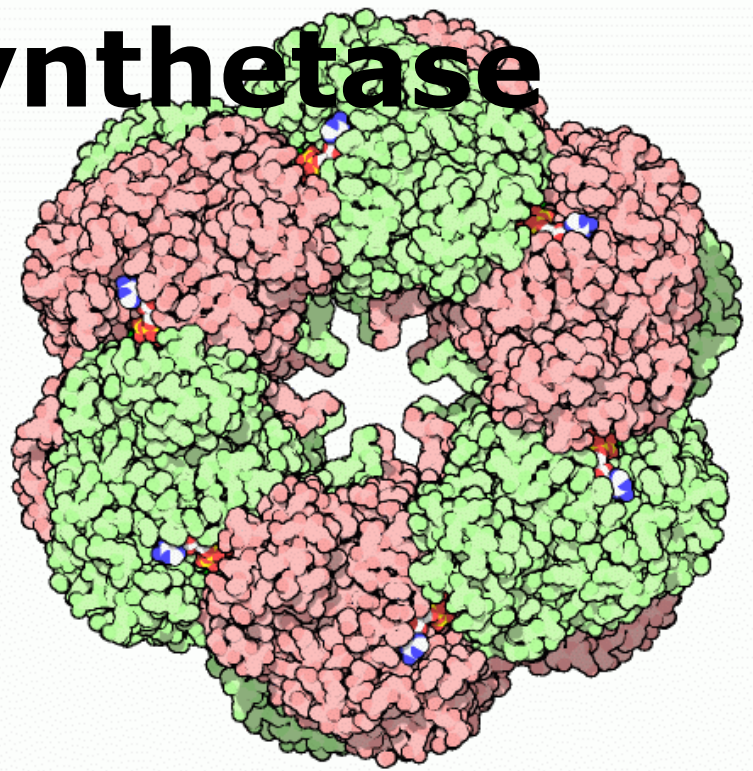
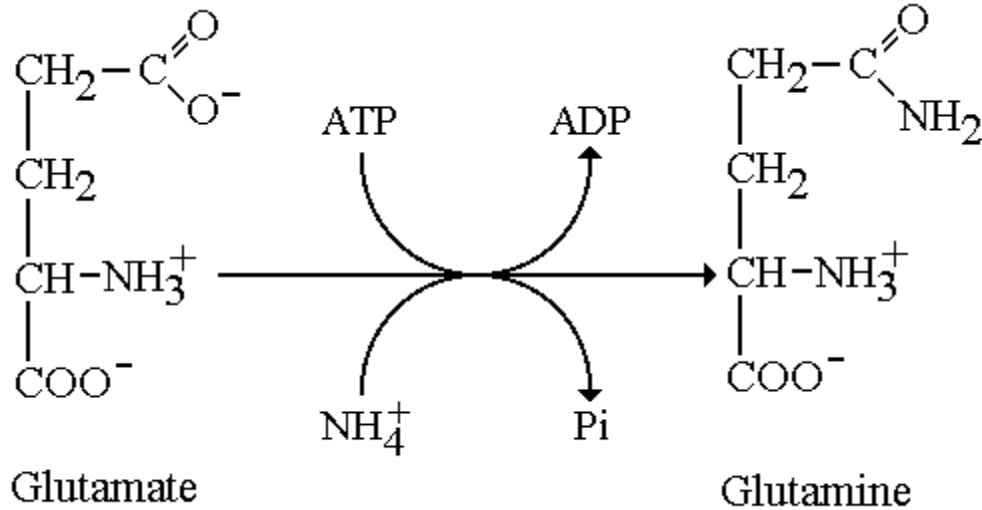
[andrade@uni-mainz.de](mailto:andrade@uni-mainz.de)

# Mount Everest

A photograph of Mount Everest, showing its snow-covered peaks and rocky ridges against a clear blue sky. The mountain is the central focus, with its jagged, snow-dusted ridges and dark rock faces. The sky is a deep, clear blue, providing a stark contrast to the white snow and dark rock.

**Age: 60M years**

# Glutamine synthetase



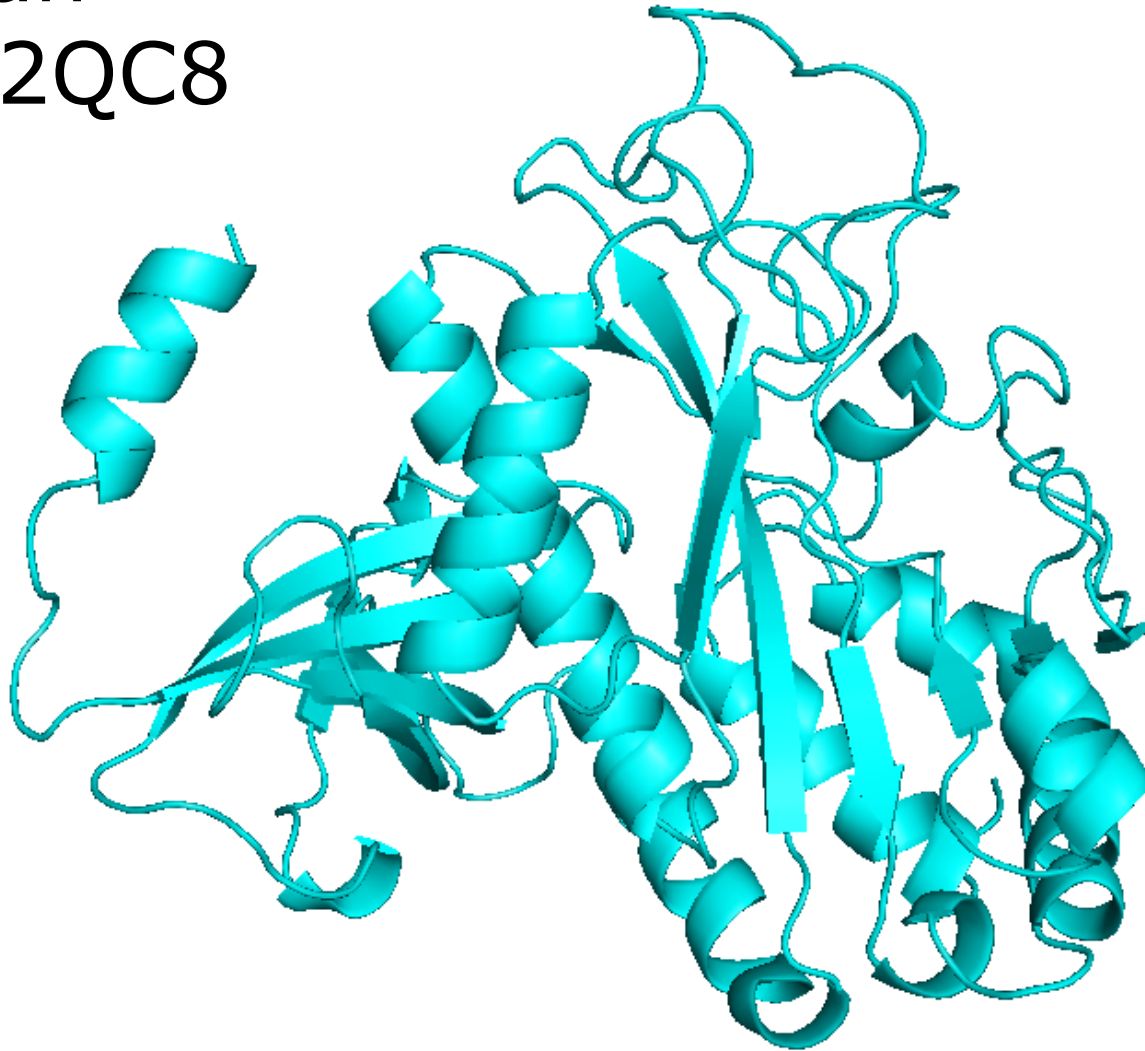
**Age: +3500M years**



# Glutamine synthetase

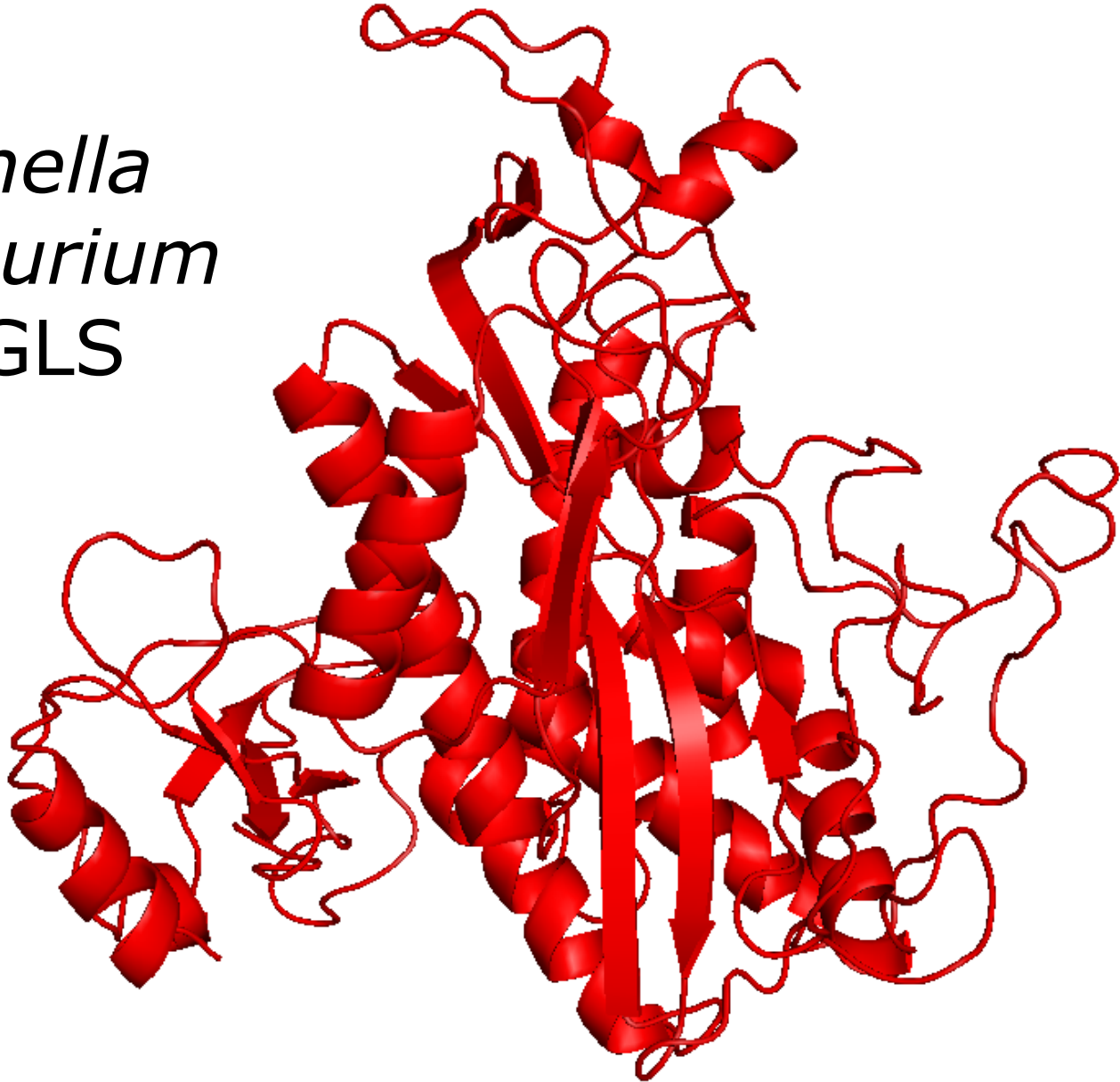
Human

PDB:2QC8

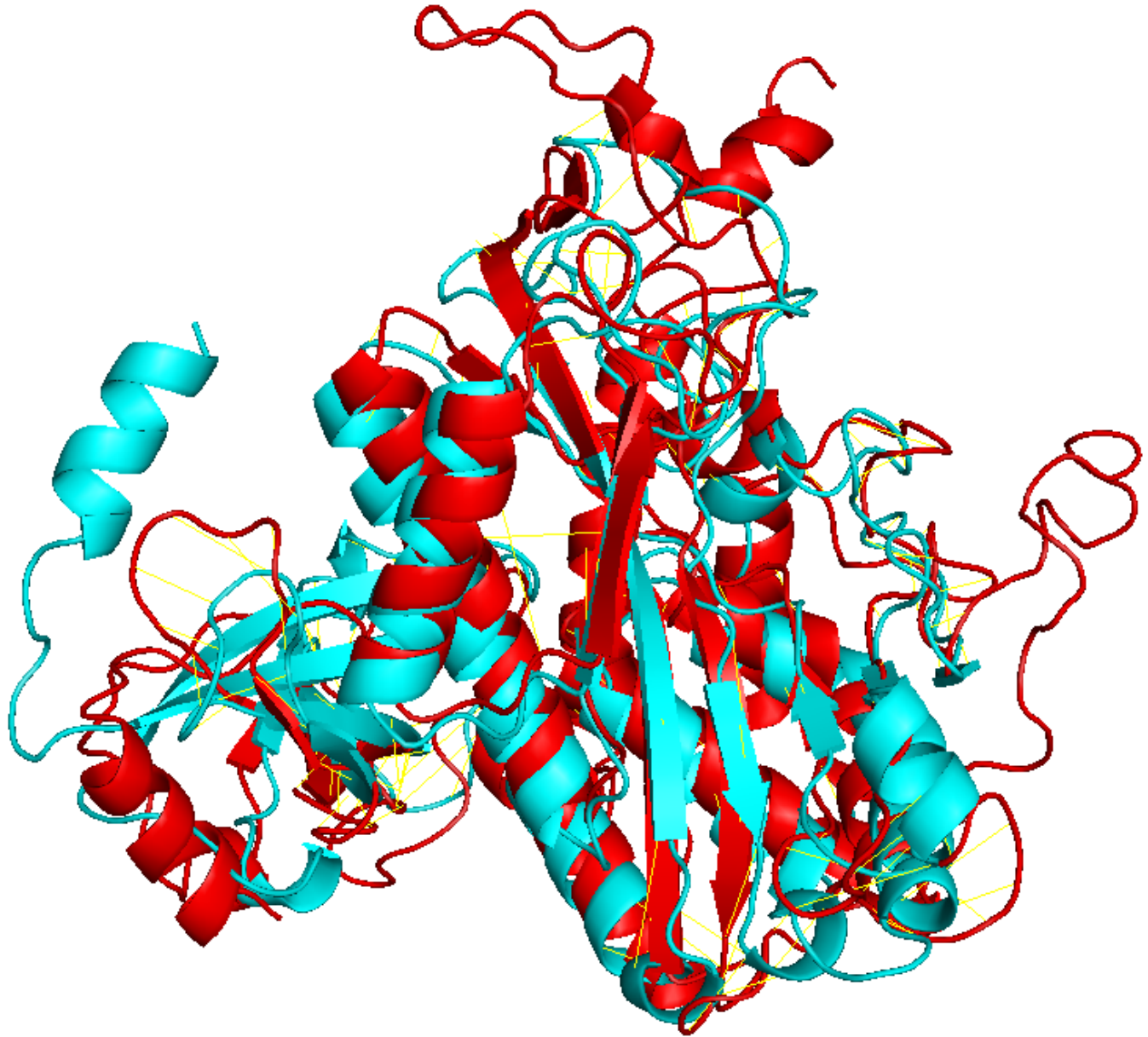


# Glutamine synthetase

*Salmonella*  
*typhimurium*  
PDB:2GLS



# Glutamine synthetase



# Time line

Earth: 4.6 By

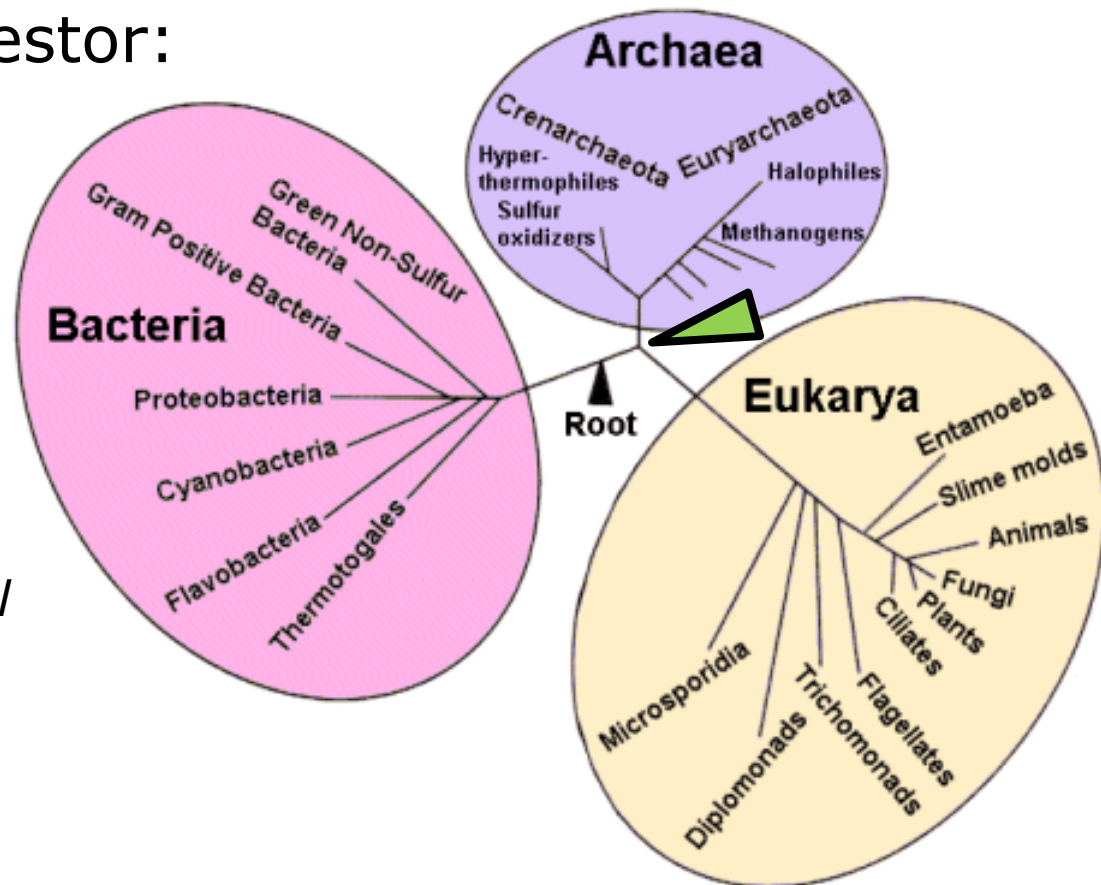
Origin of life: 3.9 By – 3.5 By

Last Common Ancestor:  
3.5 – 3.8 By

Glansdorff & Labedan  
(2008) *Biology Direct*

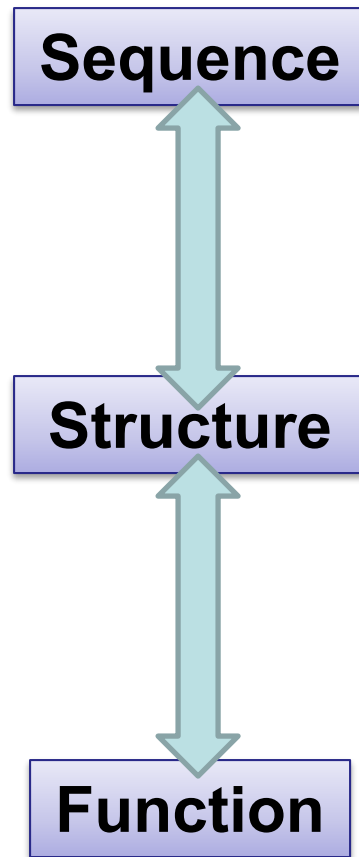
4.29 By

Sheridan *et al.* (2003)  
*Geomicrobiology Journal*



# Sequence and function

Evolutionary constraints



**MTQDELKKAVGWAALQYVQ**

**PG**

**LG**

**EK**

**DA**

**ST**

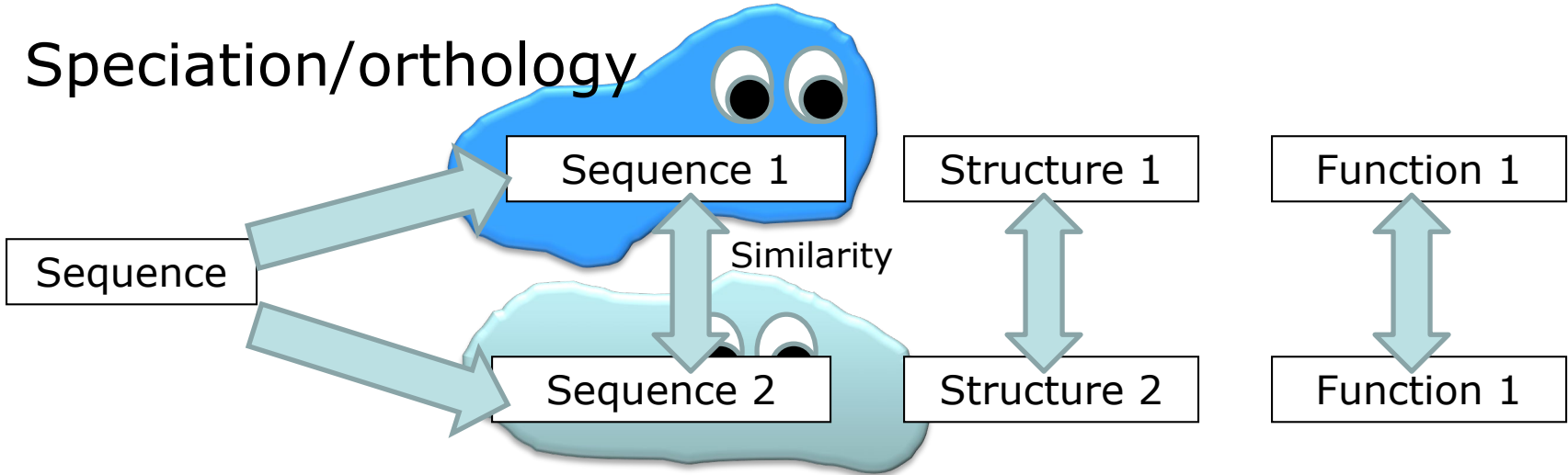




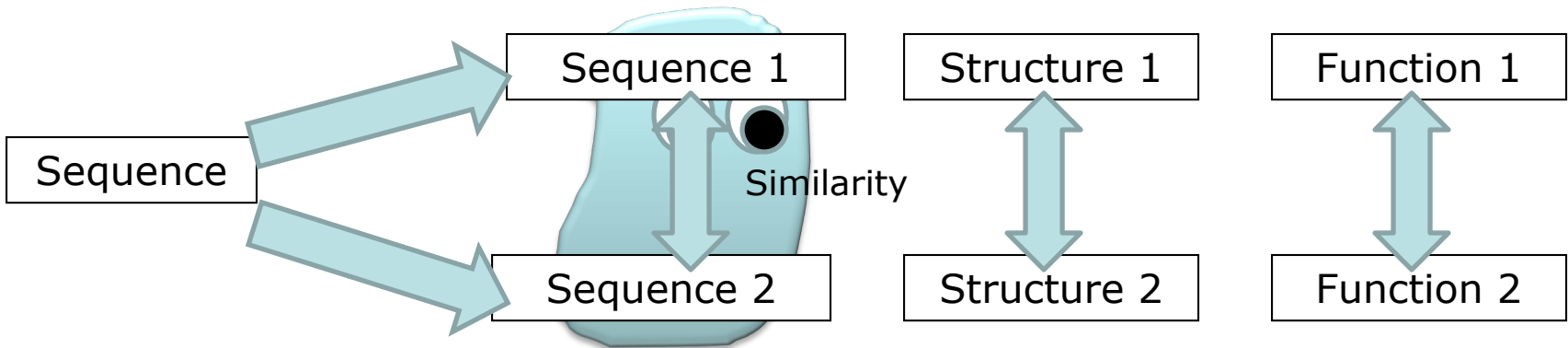
# Sequence and function

Evolutionary constraints

Speciation/orthology



Gene duplication/paralogy



# Sequence pairwise alignment

```
>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIDGTGEGLRCKTRTLTLDSEPKCVEELPEWNF DGSSTLQS
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEY TLM
GTDGHPFGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPAQWEFQIGPCE
GISMGDHLWVARFILHRVCEDFGVIATFDPKPIPGNWNAGCHTNFSTKAMREENGLKYIEEAIEKLSKR
HQYHIRAYDPKGGLDNARRLTGFHETSNINDFSAGVANRSASIRIPRTVGOEKKGYFEDRRPSANCDPFS
VTEALIRTCLLNETGDEPFQYKN
```

```
>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella
enterica subsp. enterica serovar Typhimurium str. LT2]
MSAEHVLTMLNEHEVKFVDLRFTDTKGKEQHVTIPAHQVNAEFFEKGKMGFDGSSIGGWKGINESDMVLMP
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIAKRAEDYLRATGIADTVLFGPEPEFFLFDDIR
FGASISGSHVAIDDIEGAWNSSTKYEGGNKGHRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH
HHEVATAGQNEVATRFNTMTKKADEIQIYKYVVHNVHRFGKTATFMPKPMFGDNGSGMHCHMSLAKNGT
NLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA
SPKARRIEVRFDPDPAANPYLCFAALLMAGLDGIKNKIHPGEAMDKNLYDLPPEEAKEIPQVAGSLEEALN
ALDLDFLKAAGGVFTDEAIDAYIALRREEDDRVRMTPHPVEFELYYSV
```

# Sequence pairwise alignment

## BLAST (Altschul et al, 1990)

>lcl|39919 unnamed protein product  
Length=469

Score = 70.5 bits (171), Expect = 1e-17, Method: Compositional matrix adjust.  
Identities = 102/363 (28%), Positives = 138/363 (38%), Gaps = 96/363 (26%)

```
Query 62 FDGSSSTLQSEGSN-SDMYLVPAA--MFRDPFRKDPNKLVLCEVFK-----YNRRP---- 108
          FDGSS  +G N SDM L+P A      DPF D  ++ C++ +      Y+R P
Sbjct 50 FDGSSIGGWKGINESDMVLMPDASTAVIDPFFADSTLIIRCDILEPGTLOGYDRDPRSIA 109

Query 109 --AETNLRHTCKRIMDMVSNQHPWFGMEQEYTLMGTDGHPFGWPSNGF----- 154
          AE LR T  I D V      FG E E+ L  D  FG  +G
Sbjct 110 KRAEDYLRTG--IADTV-----LFGPEPEFFLF--DDIRFGASISGSHVAIDDIEGAWN 160

Query 155 -----PGPQGPYYCGVGADRAYGRDI-----VEAHYRACLYAG 187
          PG +G Y+      D A  +DI      VEAH+      AG
Sbjct 161 SSTKYEGGNKGHRRPGVKGGYFPVPPVDSA--QDIRSEMCLVMEQMGLVVEAHHHEVATAG 218

Query 188 VKIAGTNAEVMPAQWEFQIGPCEGISMGDHLWVARFILHRVCEDFGVIATFDPKPIPG-N 246
          T  M  +      D + + ++++H V  FG ATF PKP+ G N
Sbjct 219 QNEVATRFNTMTKK-----ADEIQIYKYVVHNVAHREFGKTATFMPKPMFGDN 265

Query 247 WNGAGCHTNFSTKAMREENGLKYIEEAIEKLSKRHQYHIRAYDPKGGLDNA----- 297
          +G CH + +      +G KY      LS++  Y+I      NA
Sbjct 266 GSGMHCHMSLAKNGTNLFSGDKY-----AGLSEQALYYIGGVIKHAKAINALANPTTNSY 320

Query 298 RRLTGFHETSNINDFSAGVANRSASIRIPRTVQEKKGYPEDRRPSANCDPFSVTEALIR 357
          +RL  +E  +  +SA  NRSASIRIP V  K  E R P  +P+  AL+
Sbjct 321 KRLVPGYEAPVMLAYSAA--RNRSASIRIP-VVASPKARRIEVRFDPDPAANPYLCFAALLM 377

Query 358 TCL 360
          L
Sbjct 378 AGL 380
```

# Multiple sequence alignment

```
>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVVQAMYIWIWIDGTGEGLRCKTRTLTLDSEPKVCEELPEWNFDSSTLQS
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFKNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEYTLM
GTDGHPFGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPAQWEFQIGPCE
GISMGDHLWVARFILHRVCEDFGVIAFTDPKPIPGNWNWAGACHTNFTKAMREENGLKYIEEAIEKLSKR
HQYHIRAYDPKGGDLNARRLTGFHETSNINDFSAGVANRSASIRIPRTVVGQEKKGYPFEDRRPSANCDPFS
VTEALIRTCLLNETGDEPFQYKN
```

```
>gs_vulca gi|307594850|ref|YP_003901167.1| glutamine synthetase [Vulcanisaeta
distributa DSM 14429]
```

```
MPTRNLEIEPADLWRILKASGIKYVKFIIVDINGAPRSEIVPIDMAKDLFIDGMPFDASSIPSYSTVTKS
DFVAVYVDPRAVYVEYVQDQGVADVFTMVSDIADKPSPLDPRRVLNDALEQARSKGYEFLMGVEVEFFVIK
EDGGKPVFADPGIYFDGWNVTVQSQFMKELITAIADAGINYTKTHHEVAPSQYEVNIGATDPLRLADQIV
YFKIMAKDIARKYGLVATFMPKPFVWVNGSGAHTHISVWVDGKNLQFSSTGKITEECGYAISAILSNARA
LSSFVAPLVNSYKRLVPHYEAPTRIVWGYANRSAMIRIPQYKMRINRIEYRHPDPSMNPYLAFTAIKTM
IRGLEEKKEPPPTTEEVAYELANALETTPATLEDTLKELSKSFLATELPSSELVNAYIKIKQNEWEDYLTNV
GPWEKTWNIITQWEYNKYLVTA
```

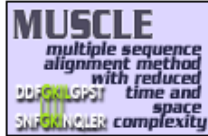
```
>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella
enterica subsp. enterica serovar Typhimurium str. LT2]
```

```
MSAEHVLTMLNEHEVKFVDLRFDTKQKQHVTPAHQVNAEFFEKGKMGFDGSSIGWKGINESDMVLMP
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDRPSIAKRAEDYLRTGIADTVLFGPEPEFFLFDDIR
FGASISGSHVAIDDIAGAWNSSTKYEGGNKGRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH
HHEVATAGQNEVATRFNTMTKKADEIQIYKYVVHNVHRFGKTATFMPKPMFGDNGSGMHCHMSLAKNGT
NLFSGDKYAGLSEQALYYIGGVKHKAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA
SPKARRIEVRFDPAAANPYLCFAALLMAGLDGKIKNIHPGEAMDKNLYDLPPPEEAKEIPQVAGSLEEALN
ALDLDREFLKGAGVFTDEAIDAYIALRREEDDRVRMT'PHPVEFELYYSV
```

```
>gs_yeast gi|330443748|ref|NP_015360.2| Gln1p [Saccharomyces cerevisiae S288c]
```


```
MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGNLRSKGRITLKKRITSIDQLPEWNFDSSTNQAPGH
DSDIYLPVAYYPDPFRRGDNIIVLAACYNNDGTPNKNFNRHEAAKLFAAHKDEEIVFGLEQEYTLFDMY
DDVYGWPKGGYPAPQGPYYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPQSQWVQVGPCTGID
MGDQLWMARYFLHRVAEEFQGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPGGMKYIEQAIEKLSKRHAE
HIKLYGSDNDMRLTGRHETASMTAFSSGVANRGSSIRIPRSVAKEGYGYFEDRRPASNIDPYLVGTGIMCE
TVCGAIDNADMTKEFERESS
```





- [Help](#)
- [MUSCLE website](#)
- [Jalview](#)
- [Programmatic Access](#)
- [Download](#)

- [Related Applications](#)
  - [Pairwise Sequence Alignment](#)
  - [Multiple Sequence Alignment](#)
  - [Phylogeny](#)

**MUSCLE related literature** 

Search for MUSCLE related literature in Medline... [more](#)

EBI > Tools > Multiple Sequence Alignment > MUSCLE

## MUSCLE - Multiple Sequence Alignment

MUSCLE stands for **M**ultiple **S**equence **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

**Internet Explorer users: If button presses (including copy/paste operations) don't appear to work please try enabling Compatibility View.**

### Use this tool

#### STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Or upload a file:  No file chosen

#### STEP 2 - Set your Parameters

OUTPUT FORMAT:

*The default settings will fulfill the needs of most users and, for that reason, are not visible.*

*(Click here, if you want to view or change the default settings.)*

#### STEP 3 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

<http://www.ebi.ac.uk/Tools/msa/muscle/>

>gs\_human gi|74271837|ref|NP\_001028216.1| glutamine synthetase [Homo sapiens]  
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIWDTGEGLRCKTRTLTLDSEPKCVEELPEW  
N-FDGSSTLQSEGSNSD---MYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPA-ETNLRH  
TCKRIMDMVSNQH----PWFMEQEYTLTGMT-----DGHPIFGW-----  
-PSNGFPGPQGP--YYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPA-QWEFQ  
IGPCEGISMGDHLWVARFILHRVCEDFGVIATFDPKPIPGNWNAGGCHTNFSTKAMREEN  
GLKYIEEAIEKLSKRHQYHIRAYDPKGG-----LDNARRLTGFHETSININDFSAGV  
ANRSASIRIPRTVQGEKKGYFEDRRPSANCDPFSVTEALIRT-CLLNETGDEP-----  
-----  
-----FQYKN-----

>gs\_yeast gi|330443748|ref|NP\_015360.2| Gln1p [Saccharomyces cerevisiae S288c]  
--MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGN-LRSKGRTLKKRITSIDQLPEW  
N-FDGSSTNQAPGHSD---IYLKPVAYYDPFRRGDNIVVLAACYNNDGTPN-KFNHRH  
EAAKLFAAHKDEE----IWFGLEQEYTLFDM-----YDDVYGW-----  
-PKGYPAPQGP--YYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPA-QWEFQ  
VGPCTGIDMGDQLWMARYFLHRVAEEFGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPG  
GMKYIEQAIEKLSKRHAHEHIKLYG-----SDNDMRLTGRHETASMTAFSSGV  
ANRGSSIRIPRSVAKEGYGYFEDRRPASNIDPYLVGTGIMCETVCGAIDNADMT-----  
-----  
-----KEFERESS-----

>gs\_vulca gi|307594850|ref|YP\_003901167.1| glutamine synthetase [Vulcanisaeta distributa DSM 14429]  
MPTRNLEIEPADLWRI---LKASGIKYVKFIIIVDINGA---PRSEIVPIDMAK-DLFDIG  
MPFDASSIPSYSTVNKSDVFVAYVDPRAVYVEYWQDGKQVADVFTMVSDIADKPS-PLDPRR  
VLNDALEQARSKGYE--FLMGVEVEFFVIKE-----  
--DGGKPVFADPGIYFDGWNVTV--QSQFMKELITAIADAGINYTKTHHEVAPS-QYEVN  
IGATDPLRLADQIVYFKIMAKDIARKYGLVATFMPKPFWGV-NGSGAHTHIS---VWKDG  
KNLF-QSSTGKITEECGYAISAILSARNALSSFVAPLVNSYKRLVPHYEAPTRIVW--GY  
ANRSAMIRIPQ--YKMRINRIEYRHPDPSMNPYLAFTAI IKTMIRGLEEKKEPPPTEEV  
AYELA--NALETP---ATLEDTLK--ELSKSFLATE--LPSELVNAYIKIKQNEWEDYLT  
NVGPWEKTWNIITQWEYNKYLVT

>gs\_salmonella gi|16767272|ref|NP\_462887.1| glutamine synthetase [Salmonella enterica]  
-----MSAEHVLTM---LNEHEVKFVDLRFTDTK GK---EQHVTIPAHQVNAEFFEEG  
KMFDSIGGWKGINESDMVLMPPDASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRS  
IAKRAEDYL RATGIADTVLFGPEPEFFLFDDIRFGASISGSHVAIDIEGAWNSSTKYEG  
GNKGHRPGVKGG--YFPVPPVDS--AQDIRSEMCLVMEQMGLVVEAHHHEVATAGQNEVA  
TRFNTMTKKADEIQIYKYVVHNVVHRFGKTATFMPKPMFGD-NGSGMHCHMS---LAKNG  
TNLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAY--SA  
RNRASIRIPV-VASPKARRIEVRFDPANPYLCFAALLMAGLDGIKNIHPGEAMDKN  
LYDLPPEEAKEIPQVAGSLEEALNALDLDFL KAGGVFTDEAIDAYIALRREEDDRVRM  
TPHP-----VEFELYYSV-



ClustalW, JalView

# Determination of protein structure

X-ray crystallography (111K in PDB)

- need crystals

Nuclear Magnetic Resonance (NMR)  
(10.4K)

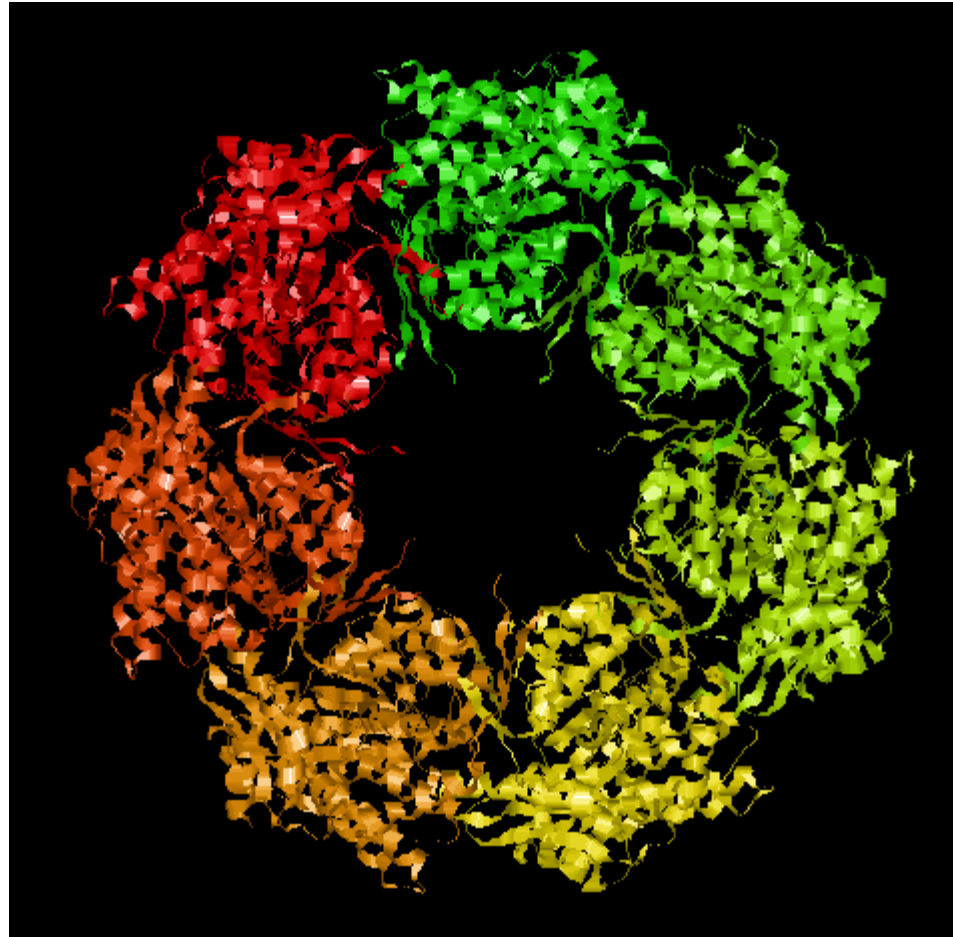
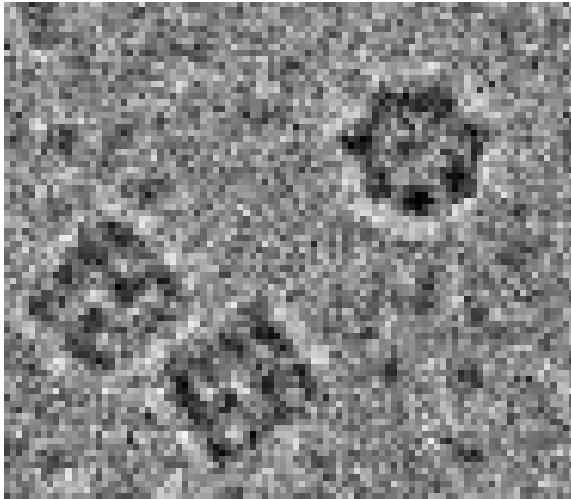
- proteins in solution
- lower size limit (600 aa)

Electron microscopy (1.2K)

- Low resolution (>5Å)

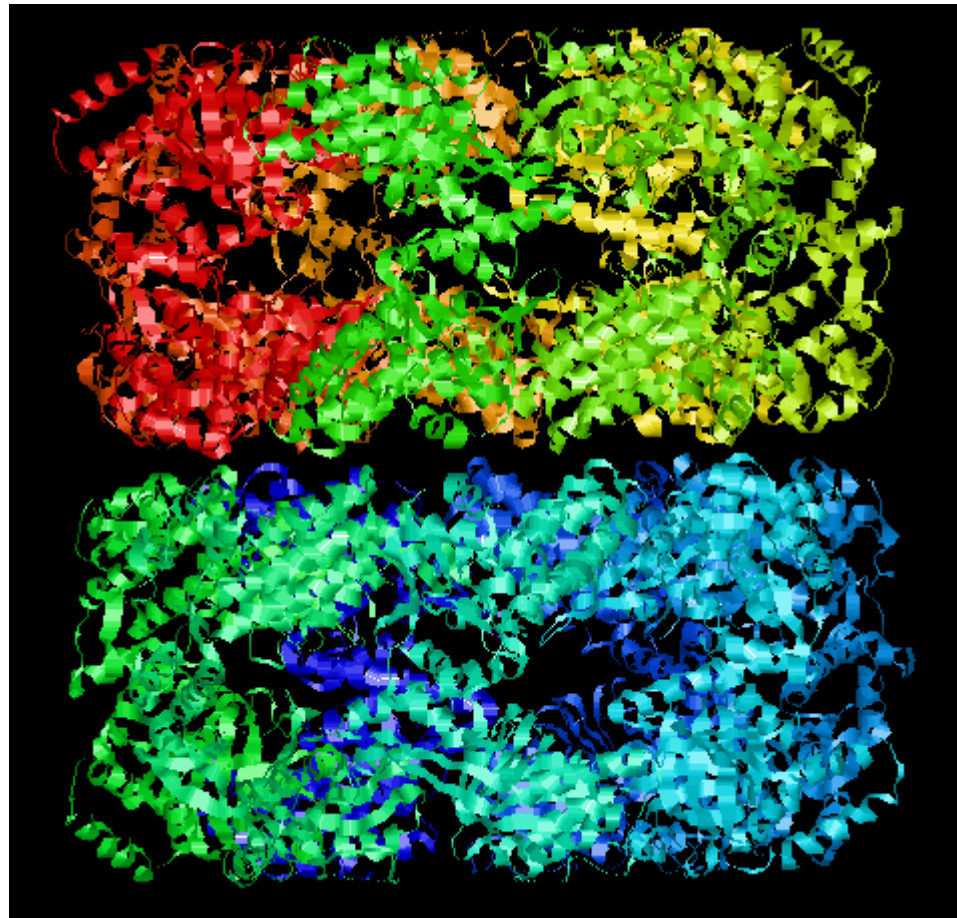
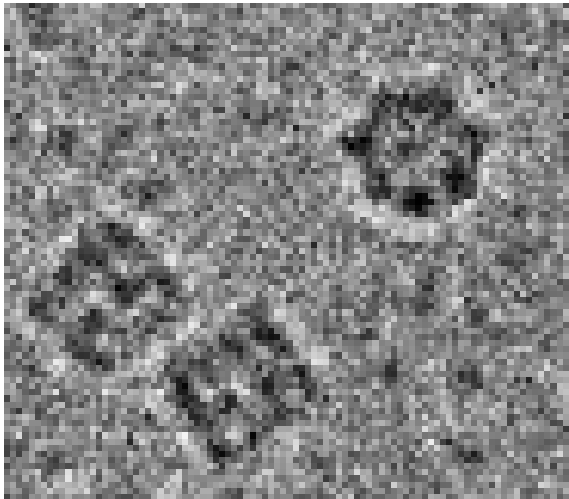


# Determination of protein structure



resolution 2.4 Å

# Determination of protein structure



resolution 2.4 Å

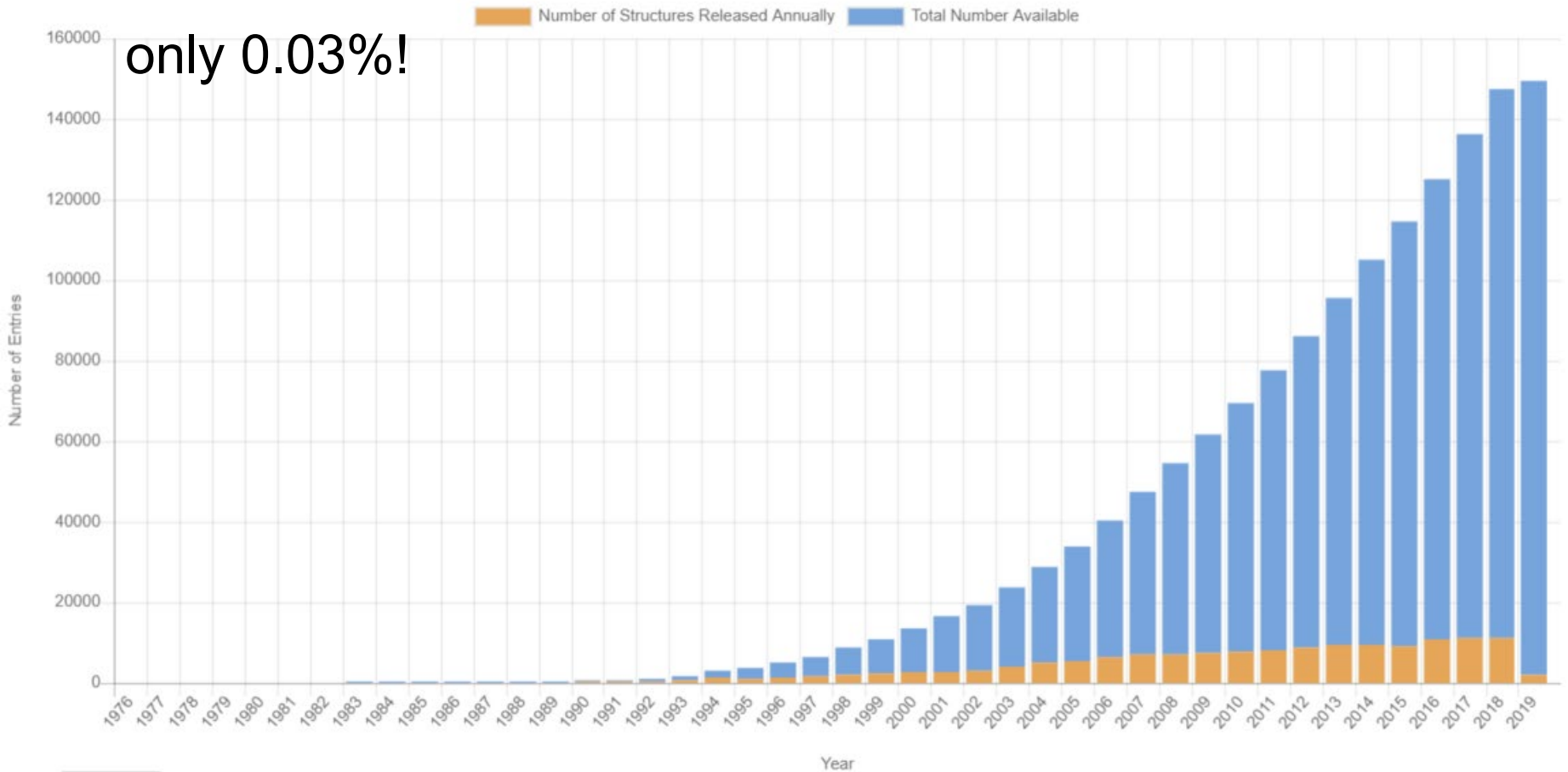
# Structural genomics

Currently: 150K protein 3D structures

from around 46.4K sequences in UniProt (how do I know?)

146M sequences in UniProt

only 0.03%!

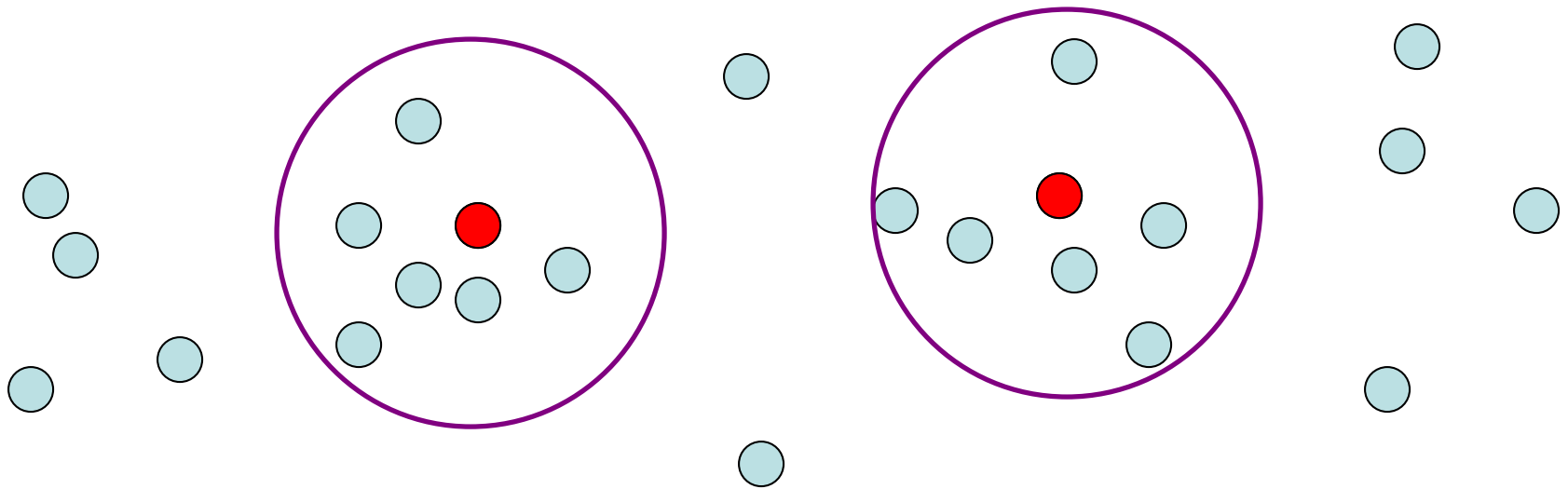


# Structural genomics

Currently: 150K protein 3D structures  
from around 46.4K sequences in UniProt (how do I know?)

146M sequences in UniProt

only 0.03%!

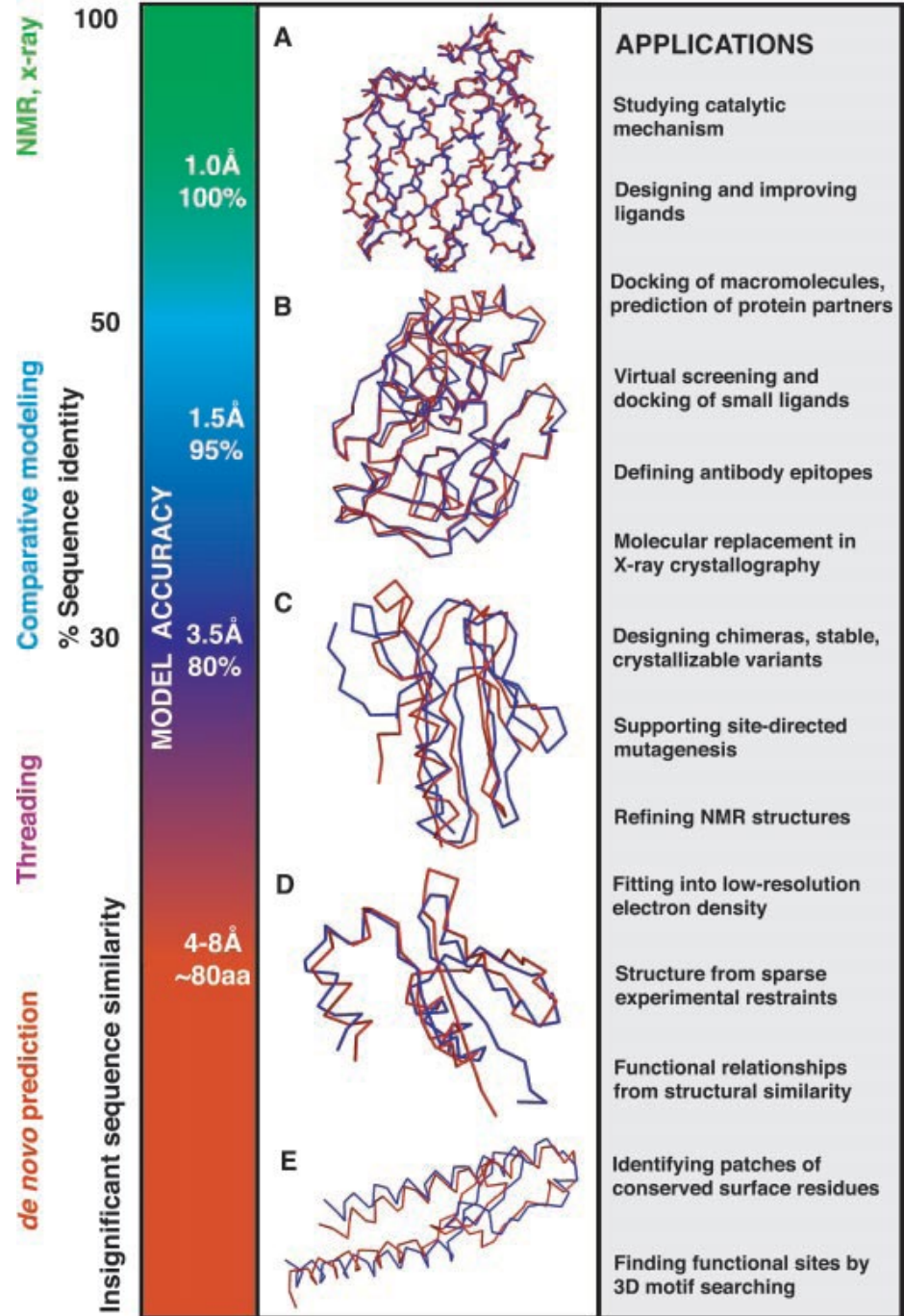


50% sequences covered (25% in 1995)



# Relation between sequence identity and accuracy/applications

From:  
Baker and Sali (2001)  
*Science*



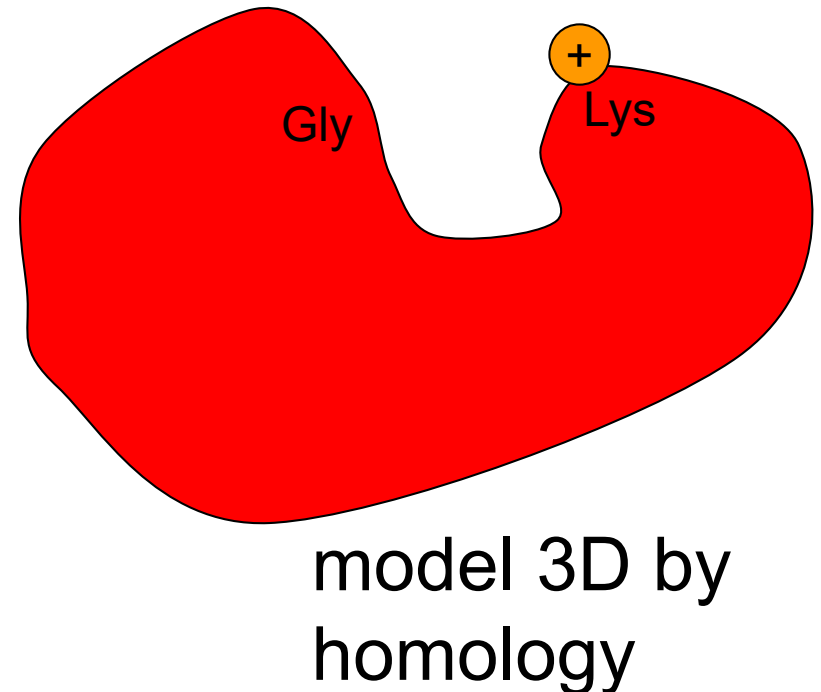
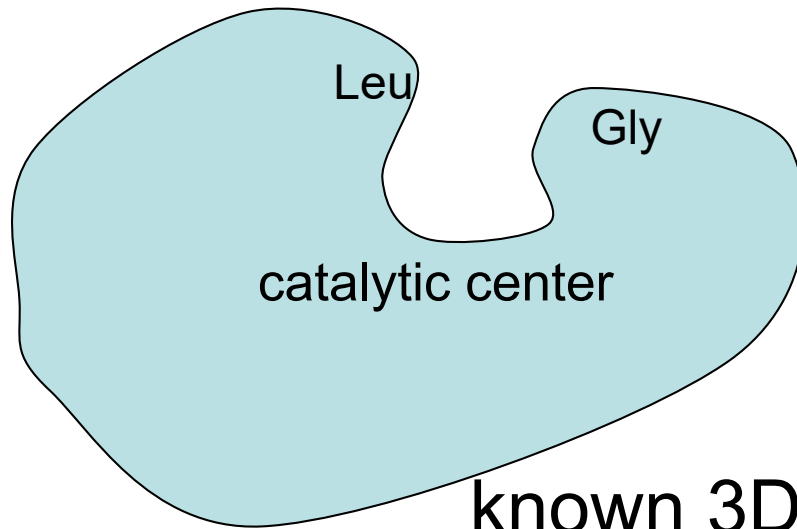
# Homology modelling

## Applications: target design

Query sequence



similar to



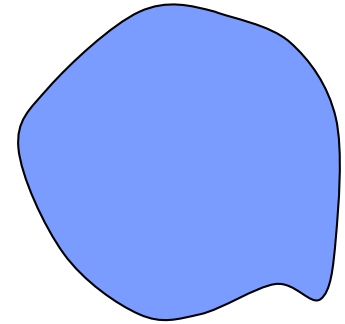
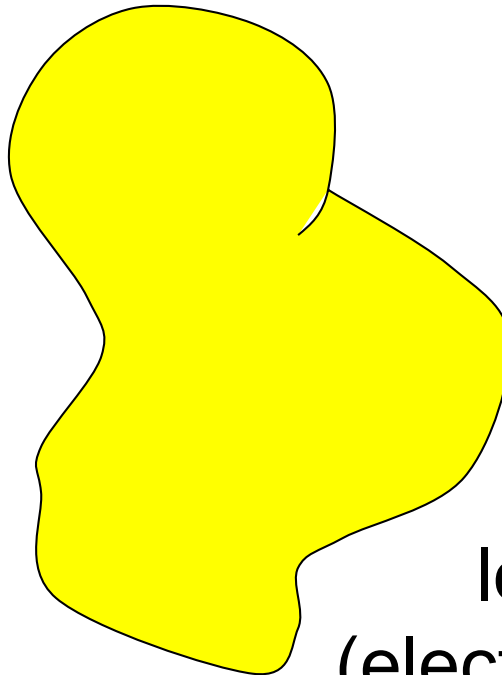
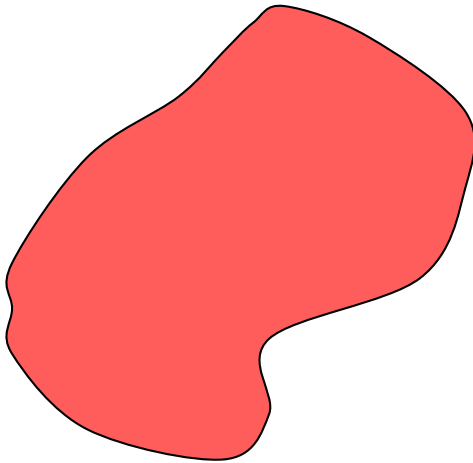
# Homology modelling

## Applications: fit to low res 3D

Query sequence 1



Query sequence 2



low resolution 3D  
(electron microscopy)

# Homology modelling

# GenTHREADER



David Jones [http://bioinf.cs.ucl.ac.uk/psipred\\_new/](http://bioinf.cs.ucl.ac.uk/psipred_new/)  
Input sequence or MSA

**Choose Prediction Methods**

<input type="checkbox"/> PSIPRED v3.3 (Predict Secondary Structure)	<input type="checkbox"/> DISOPRED3 & DISOPRED2 (Disorder Prediction)
<input checked="" type="checkbox"/> pGenTHREADER (Profile Based Fold Recognition)	<input type="checkbox"/> MEMSAT3 & MEMSAT-SVM (Membrane Helix Prediction)
<input type="checkbox"/> BioSerf v2.0 (Automated Homology Modelling)	<input type="checkbox"/> DomPred (Protein Domain Prediction)
<input type="checkbox"/> FFPred v2.0 (Eukaryotic Function Prediction)	<input type="checkbox"/> GenTHREADER (Rapid Fold Recognition)
<input type="checkbox"/> MEMPACK (SVM Prediction of TM Topology and Helix Packing)	<input type="checkbox"/> pDomTHREADER (Fold Domain Recognition)
<input type="checkbox"/> DomSerf v2.0 (Automated Domain Modelling by Homology)	

[Help...](#)

**Input Sequence (Single sequence or Multiple Sequence alignments; as raw sequence or fasta format)**

Typically 30 minutes, up to two hours

GenTHREADER Jones (1999) *J Mol Biol*

# Homology modelling Phyre



Mike Sternberg <http://www.sbg.bio.ic.ac.uk/phyre2/>

Kelley et al (2000) *J Mol Biol*  
Kelley et al (2015)  
*Nature Protocols*

The screenshot shows the Phyre2 web interface. At the top, the logo 'Phyre2' is displayed in large, 3D-style letters. Below the logo, the text 'Protein Homology/analogY Recognition Engine V 2.0' is visible. To the right of the logo, there is a 'Subscribe to Phyre at Google Groups' section with an email input field and a 'Subscribe' button. Below this, there is a link to 'Visit Phyre at Google Groups'. A row of icons (calendar, magnifying glass, question mark, envelope, books) is positioned below the subscription section. A link for 'What's New in Phyre2' is also present. The main search area contains several input fields: 'E-mail Address', 'Optional Job description', and 'Amino Acid Sequence' (with an information icon). At the bottom, there is a 'Modelling Mode' section with radio buttons for 'Normal' (selected) and 'Intensive'. Two buttons, 'Phyre Search' and 'Reset', are located at the bottom right of the search area.

Processing time can be hours

# Homology modelling

## Static solutions

Datasets of precomputed models /  
computations

Not flexible

Variable coverage

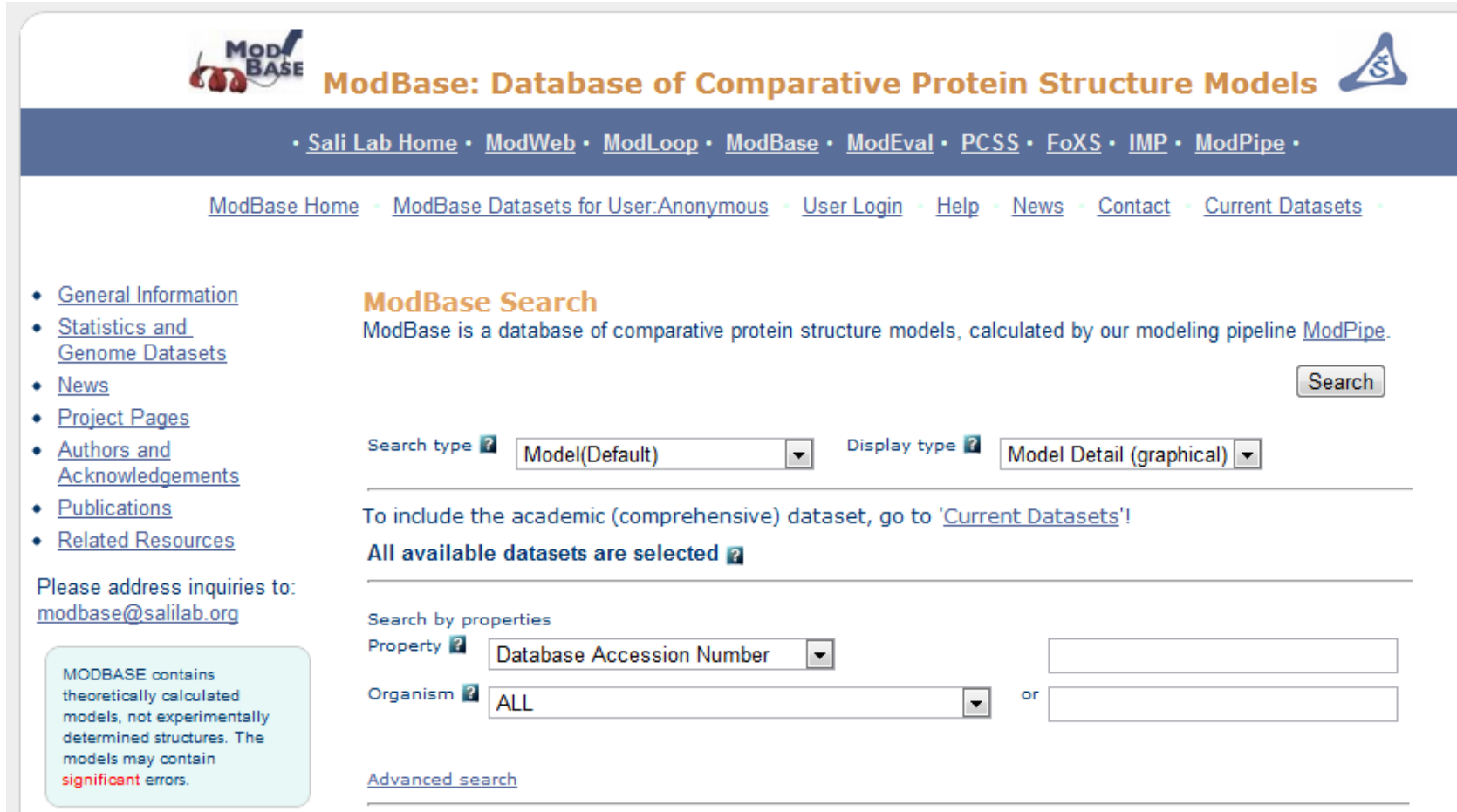
But you don't have to wait

# Homology modelling

# MODbase

Andrej Sali

<http://modbase.compbio.ucsf.edu/>



The screenshot shows the MODbase website interface. At the top, there is a logo for MODbase and the text "ModBase: Database of Comparative Protein Structure Models". Below this is a navigation bar with links to "Sali Lab Home", "ModWeb", "ModLoop", "ModBase", "ModEval", "PCSS", "FoXS", "IMP", and "ModPipe". A secondary navigation bar contains links for "ModBase Home", "ModBase Datasets for User:Anonymous", "User Login", "Help", "News", "Contact", and "Current Datasets".

On the left side, there is a vertical menu with links to "General Information", "Statistics and Genome Datasets", "News", "Project Pages", "Authors and Acknowledgements", "Publications", and "Related Resources". Below this menu, it says "Please address inquiries to: [modbase@salilab.org](mailto:modbase@salilab.org)".

The main content area features a "ModBase Search" section. It includes a "Search" button and two dropdown menus: "Search type" set to "Model(Default)" and "Display type" set to "Model Detail (graphical)". Below these is a note: "To include the academic (comprehensive) dataset, go to 'Current Datasets!'". A status message says "All available datasets are selected".

There is a "Search by properties" section with a "Property" dropdown set to "Database Accession Number" and an empty input field. Below it, the "Organism" dropdown is set to "ALL" with an empty input field, followed by the word "or" and another empty input field.

At the bottom of the search section, there is a link for "Advanced search".

A light blue box on the left contains the text: "MODBASE contains theoretically calculated models, not experimentally determined structures. The models may contain significant errors."

Pieper et al (2014) *Nucleic Acids Research*



# Homology modelling MODbase

## Sequence Overview

[Go to Model Overview](#)

Search Summary

Search Input: database\_id: sorc3\_human

Organism(s):

Homo sapiens

1 match found.

Perform Action on Selected Model(s) Check model(s), then select option

Cov	TARGET					MODEL DATA				TEMPLATE		
	Model Icon	Model/Fold Reliability	Sequence Database Link	Database Annotation	Organism	Protein Size	Modeled Segment	Size	Seq Id(%)	PDB code	PDB Segment	PDB Comment
For		<input type="checkbox"/>	<a href="#">Q5VXF9</a>	vps10 domain receptor protein sorc3 (sorc3)	<a href="#">Homo sapiens</a>	1222	198-643	446	16.00	<a href="#">1sqjA</a>	8-581	crystal structure analysis of oligoxyloglucan reducing-end-specific cellobiohydrolase (oxg-rcbh)
		<input type="checkbox"/>	<a href="#">Q5VXF9</a>	vps10 domain receptor protein sorc3 (sorc3)	<a href="#">Homo sapiens</a>	1222	798-915	118	35.00	<a href="#">1wqoA</a>	5-122	solution structure of the pkd domain from human vps10 domain-containing receptor sorc2
		<input type="checkbox"/>	<a href="#">Q5VXF9</a>	vps10 domain receptor protein sorc3 (sorc3)	<a href="#">Homo sapiens</a>	1222	198-712	515	12.00	<a href="#">1sqjA</a>	8-730	crystal structure analysis of oligoxyloglucan reducing-end-specific cellobiohydrolase (oxg-rcbh)

# Homology modelling

# Protein Model Portal



Torsten Schwede

A screenshot of the PSI | The Protein Model Portal website. The header is dark red with the text "PSI | The Protein Model Portal" in orange and white. Below the header is a navigation bar with links: "Home", "Interactive Modeling", "Quality Estimation", "Protein Modeling 101", and "More". The main content area is white and contains the text "Welcome to the Protein Model Portal (PMP)". Below this is a paragraph: "PMP gives access to various models computed by comparative modeling methods provided by different partner sites, and provides access to various interactive services for model building, and quality assessment." There is a search input field with the placeholder text "Please enter your query." and a red "Search" button. Below the search button are examples: "Examples: [UniProt AC] [UniProt ID] [RefSeq] [PDBID] [Sequence] [Free Text]".

Haas et al. (2013) *Database*

# Aquaria

Sean O'Donoghue

<http://aquaria.ws/>



The screenshot displays the Aquaria web interface for protein structure analysis. The main panel shows a 3D ribbon representation of the TET1\_HUMAN protein structure, colored by domain (green, blue, yellow). A selection box highlights a specific region labeled 'A: L(1340)'. The interface includes several informational panels:

- 3D STRUCTURE:** TET1\_HUMAN sequence aligned onto TET2 structure from PDB 4nm6-A (64% sequence identity).
- ABOUT TET1\_HUMAN:** Provides functional details such as 'FUNCTION: Dioxygenase that catalyzes the conversion of the modified genomic...', 'CATALYTIC ACTIVITY: DNA 5-methylcytosine + 2-oxoglutarat + O(2) = DNA 5-...', and 'COFACTOR: Binds 1 Fe(2+) ion per subunit. Binds 3 zinc ions per subunit.'.
- ABOUT PDB 4nm6:** Contains a reference to 'Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation.' by Hu et al., Cell (2013).
- MATCHING STRUCTURES:** A table at the bottom lists other protein structures with their sequence identity percentages (65%, 64%, 34%, 26%) and corresponding PDB IDs.

O'Donoghue et al (2015) *Nature Methods*

# Domains

Protein domains are structural units (average 160 aa) that share:

Function

Folding

Evolution

Proteins normally are multidomain (average 300 aa)



# Domains

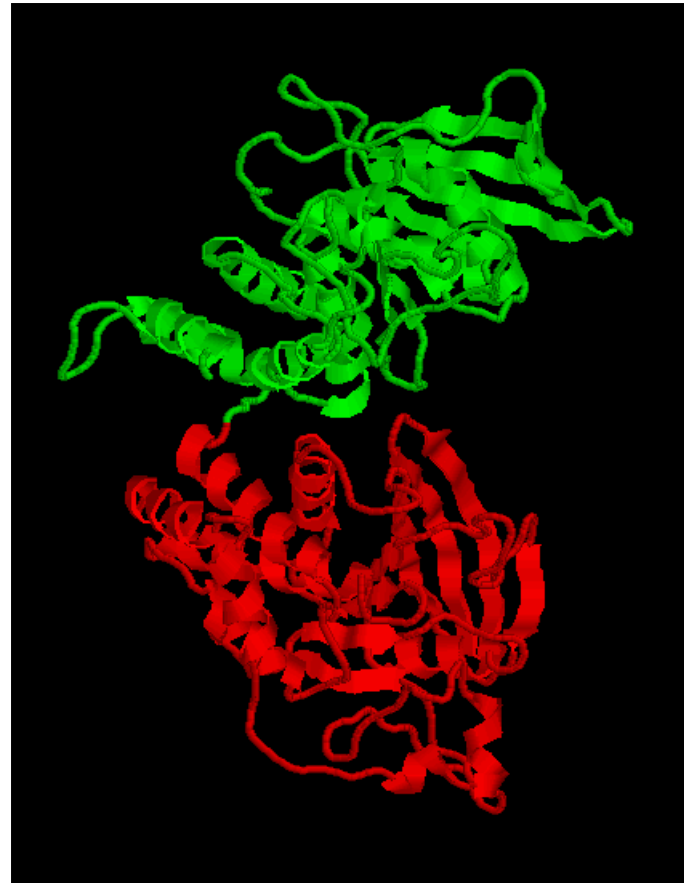
Protein domains are structural units (average 160 aa) that share:

Function


Folding


Evolution

Proteins normally are multidomain (average 300 aa)

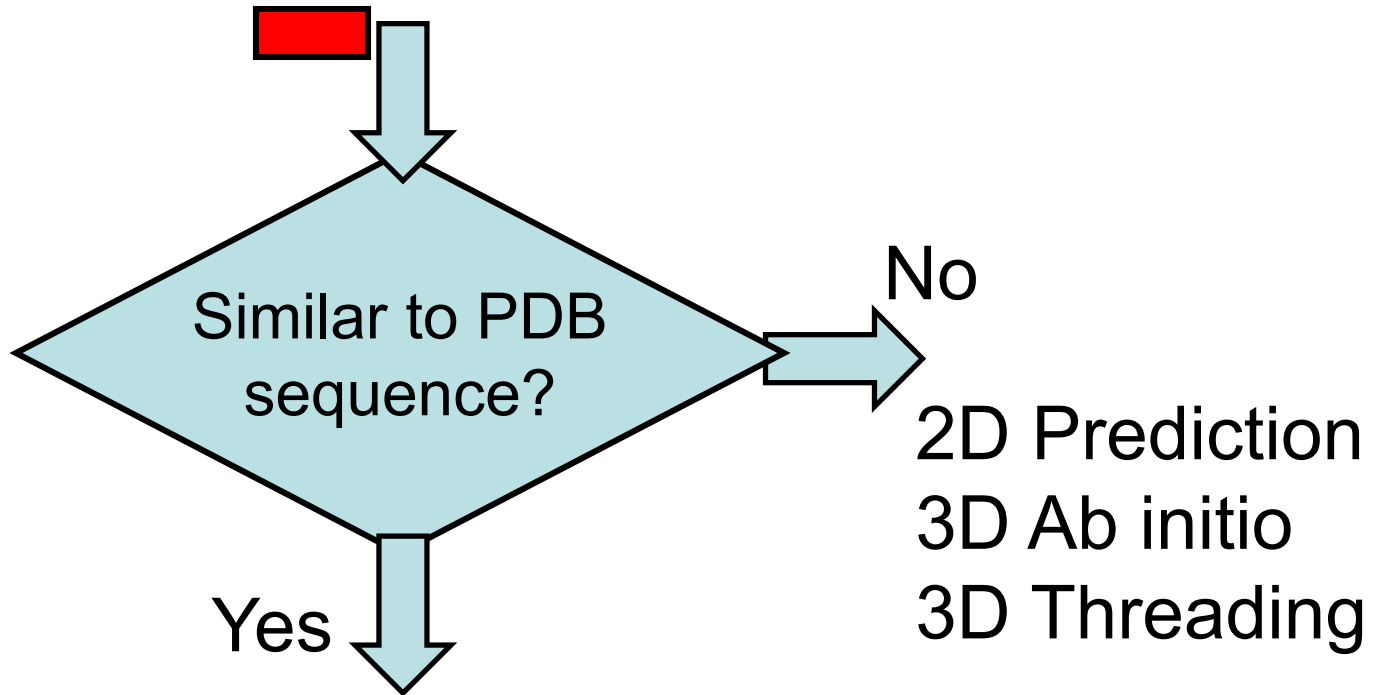


# Domains

Query Sequence 

Predict domains 

Cut



3D Modeling by homology

# 3D structure prediction

## Ab initio

Explore conformational space

Limit the number of atoms

Break the problem into fragments of sequence

Optimize hydrophobic residue burial and pairing of beta-strands

Limited success



# 3D structure prediction

# Threading

**I-Tasser:** Yang Zhang &  
Jeffrey Skolnick



Fold 66% sequences <200 aa long of low homology to PDB

Just submit your sequence and wait... (some days)

Output are predicted structures (PDB format)

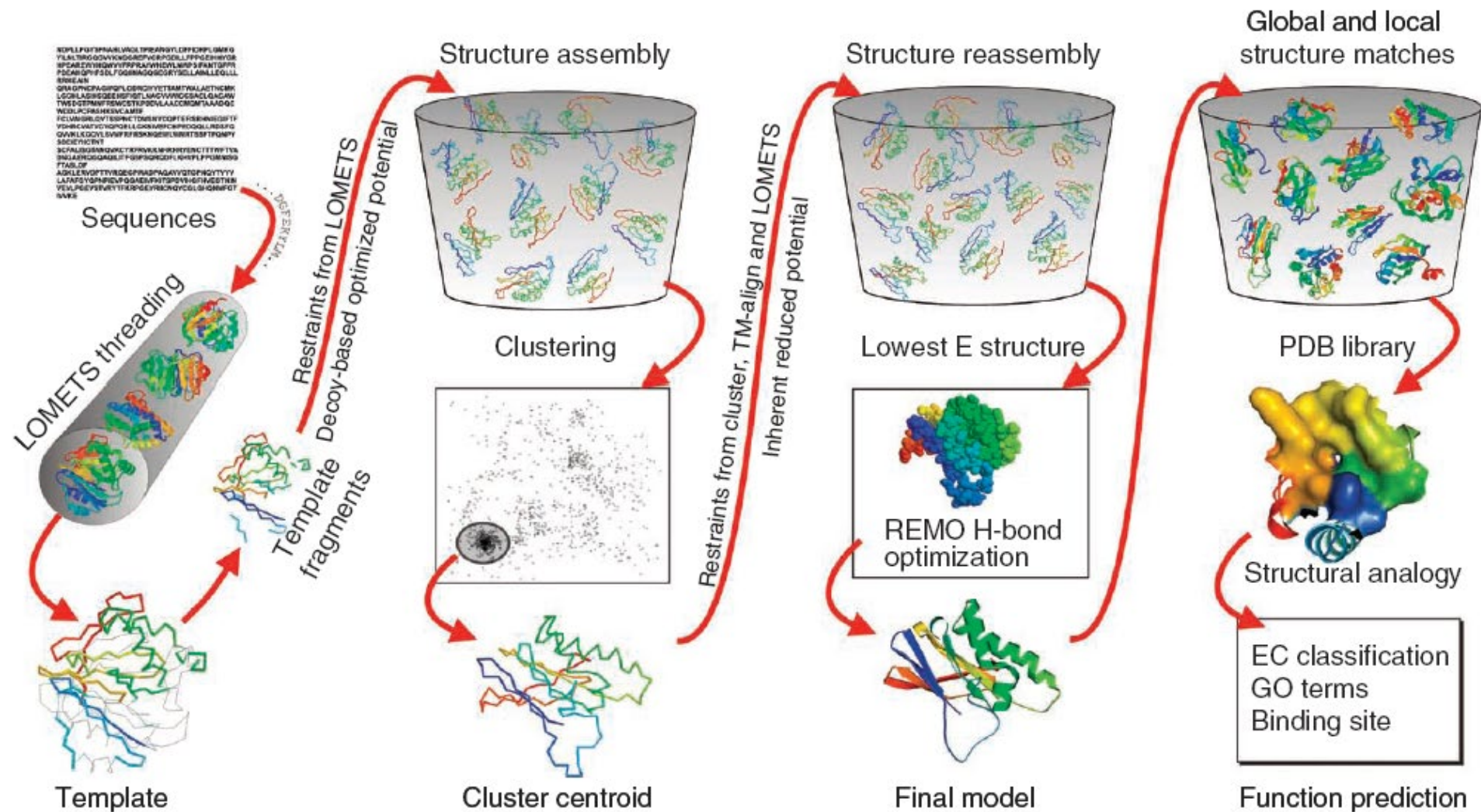
Lee and Skolnick (2008) *Biophysical Journal*

Roy et al (2010) *Nature Methods*

Yang et al (2015) *Nature Methods*

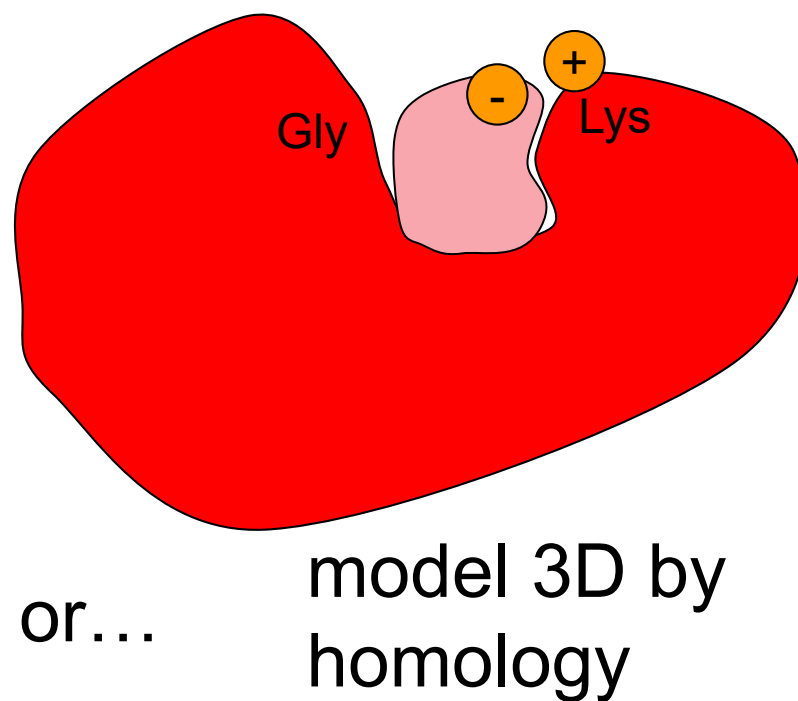
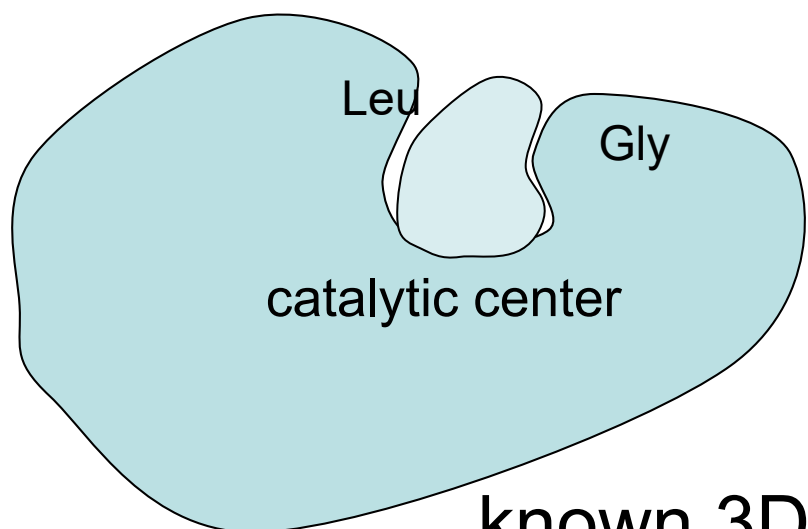
# 3D structure prediction

## I-Tasser



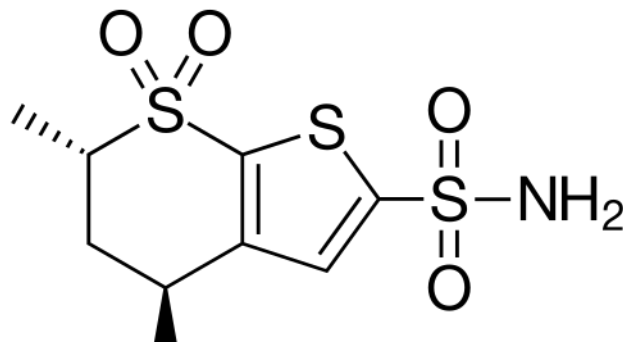
# Structure-based drug-design

Find a molecule that fits a protein



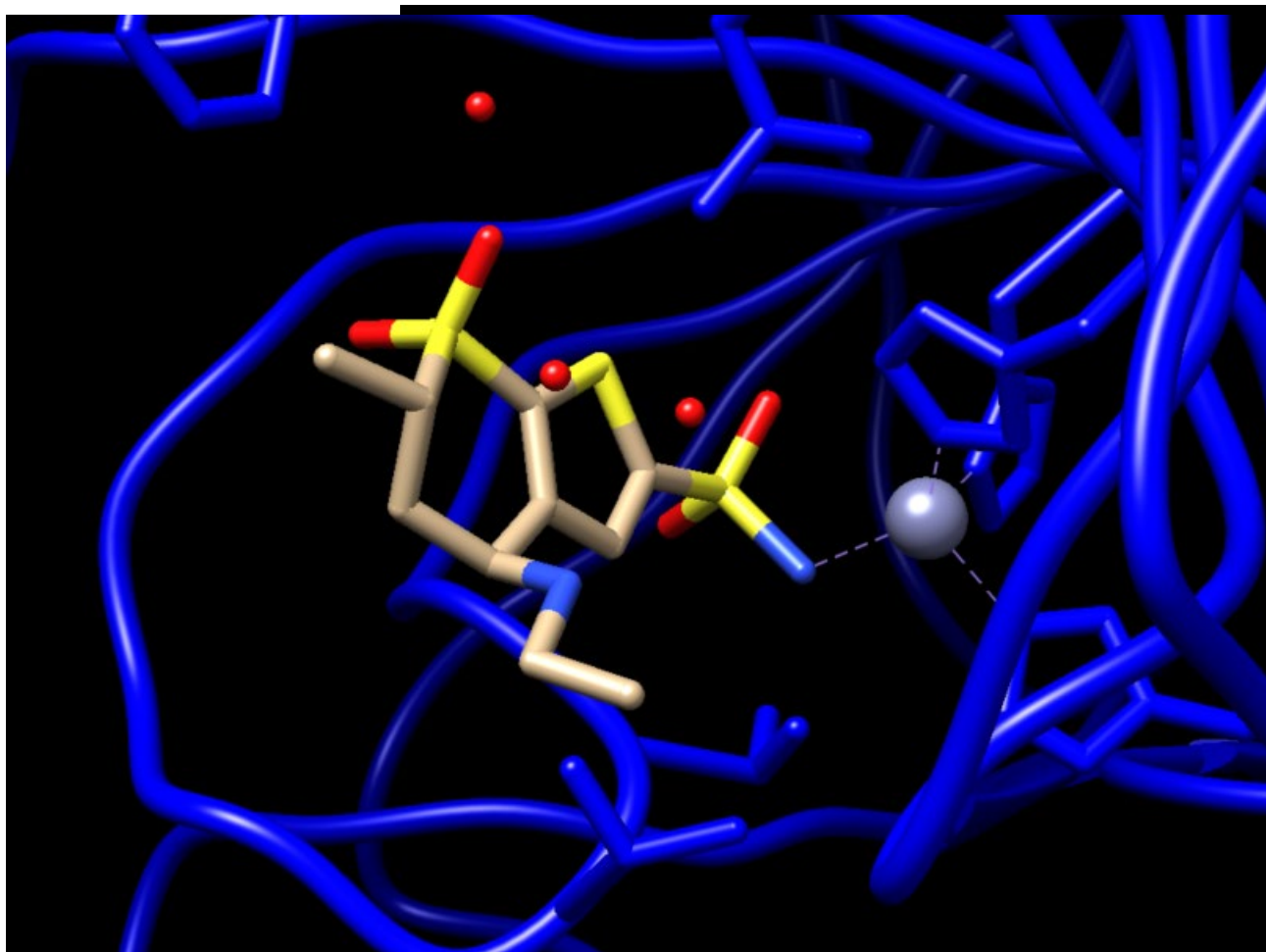
# Dorzolamide

1<sup>st</sup> structure-based drug  
design: 1995



Carbonic  
anhydrase  
inhibitor

PDB: 1CIL

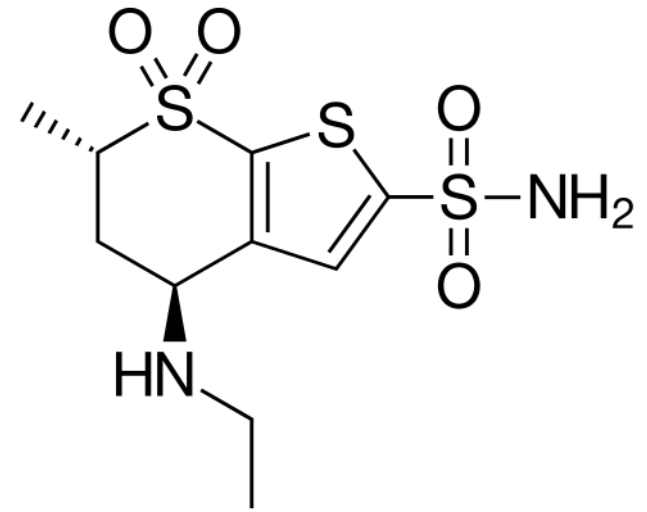


# Dorzolamide

Decreases production of aqueous humour

Reduces intraocular pressure

Eye drops to treat glaucoma and ocular hypertension

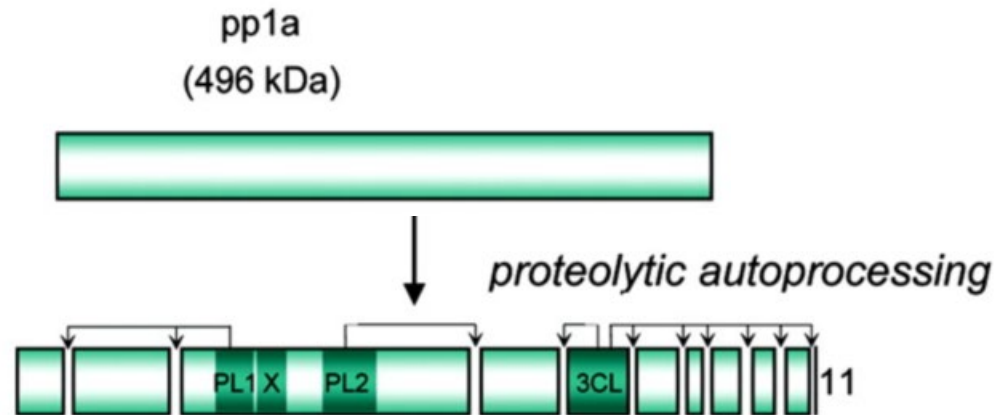




# SARS example

SARS polyprotein as a drug target

Processed by proteinase 3CLpro



# SARS example

Replicase polyprotein 1a Human SARS  
coronavirus: R1A\_CVHSA

Polyprotein:

Analyze the domains in PFAM.


Can you see the peptidase domain?

What are its coordinates (amino acid positions)?



# SARS example

Proteinase 3CLpro: activity

Cleaves at: Leu-Gln- [Ser, Ala, Gly]  
P1 - P2

Can we design a drug to block this  
proteinase?

# SARS example

We need the 3D structure of human SARS proteinase 3CLpro in complex with a target peptide.

Problem: X-ray doesn't work, cannot get crystals

We try with the proteinase 3CLpro from Porcine TGEV coronavirus (easier to culture).  
It works! PDB: 1P9U

Can we do homology modeling of the human virus protein with the porcine virus protein?

# SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

How many chains do we have? What does it mean?

What is the sequence of the peptide?

How is the peptide bound to the protease?

# SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

Select chain F. Color it red.

Select chain H. Color it yellow.

Select zone (default 5 A)

Action > Atom/bonds > show

Cysteine 144 is important!

# SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

Three domains:

domain #1: 8-99

domain #2: 100-183

domain #3: 200-300

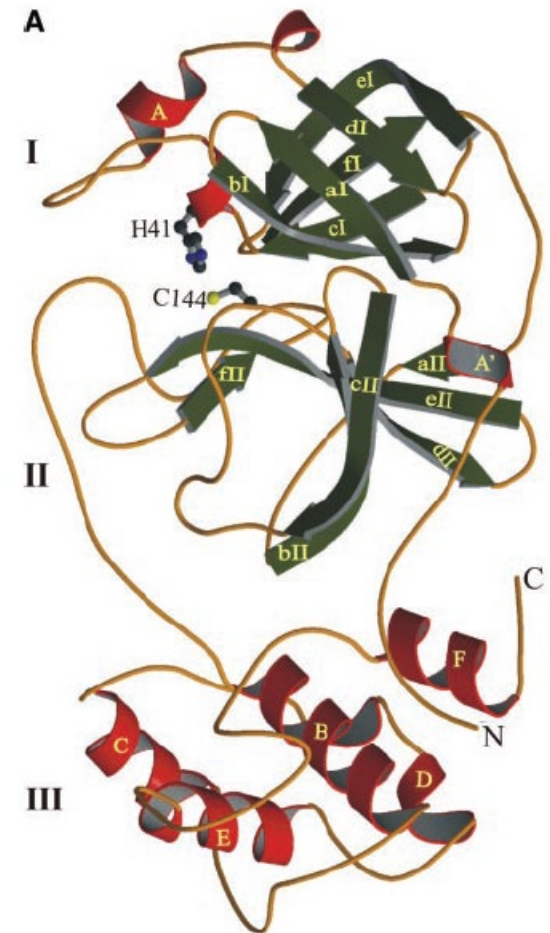
Binding site between domains #1 and #2

# SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

Catalytic diad: His 41, Cys 144



# SARS example

We need the 3CLPro from SARS.

Are they similar enough so that we could do a model?

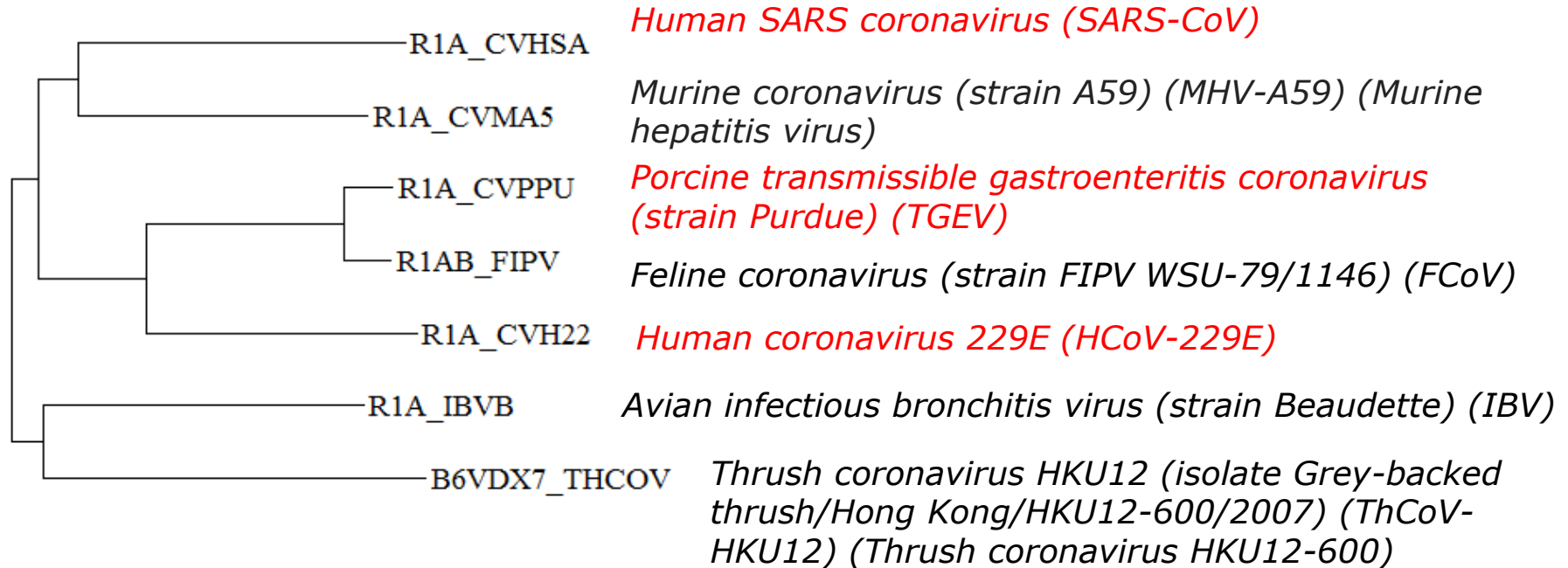
We did a multiple sequence alignment of several 3CLPro proteases TGEV, 229E, SARS and other related viruses:

From <https://cbdm.uni-mainz.de/un17/>  
Open alignment PF05409\_7seqs.txt using  
Bioedit or Jalview or Clustalw



# SARS example

Phylogenetic tree:



Is Cysteine 144 conserved? (Note that in this alignment it is at position 119).  
Anything else conserved around?

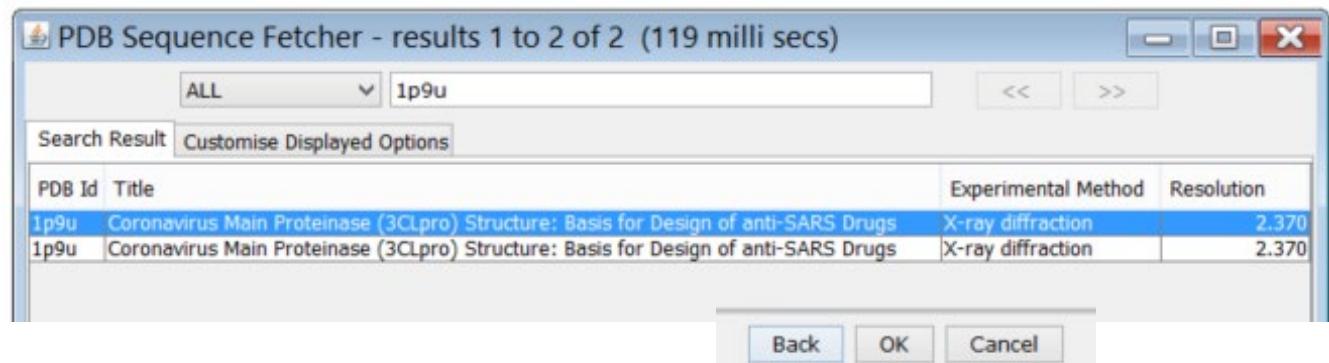
# SARS example

Compare structures of model (SARS) and TGEV protein.

Open jalview (close all the windows that appear then:  
Main Menu > Tools > Preferences > Visual:  
Tick out option "Open file")

Main menu > File > Fetch sequences > click Select  
database > click PDB

Type 1P9U, select one result and click OK



# SARS example

Now we get the file with the model. It is not in PDB.  
Find it at our course web page:

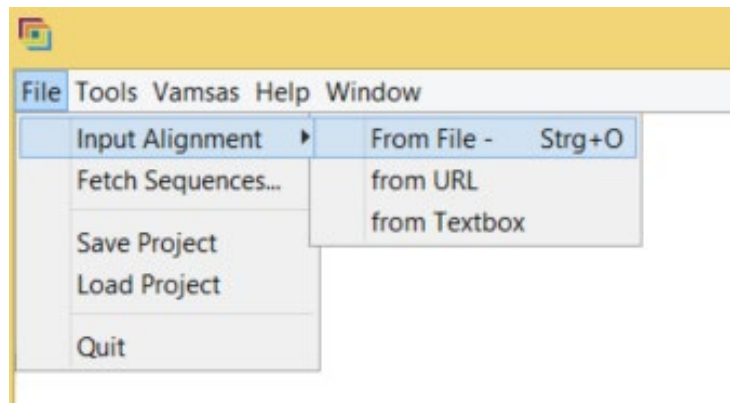
<https://cbdm.uni-mainz.de/un17/>

The name of the file is **1p9t.txt**

Right click on the link and save the file in your disk.

Then in jalview:

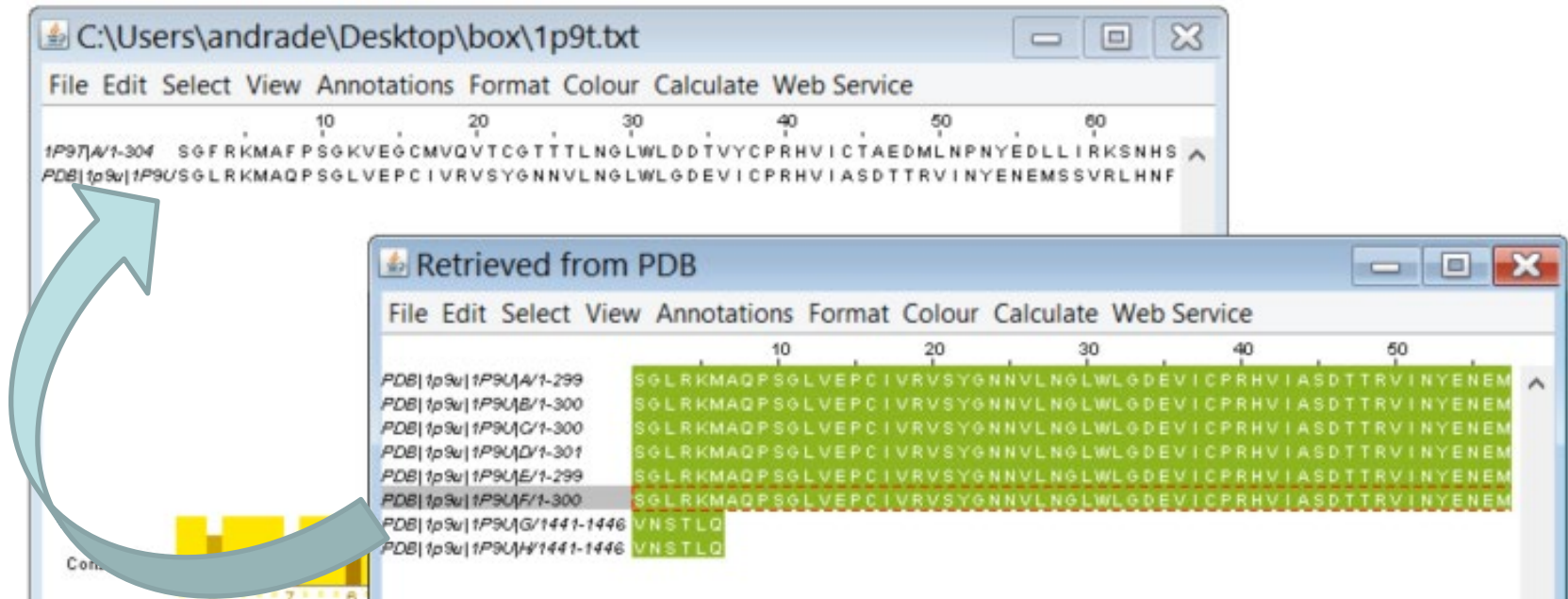
Main menu > File > Input alignment > From File



# SARS example

Next let's put together both sequences.

Copy the 1P9U|F/1-300 to the window of 1P9T



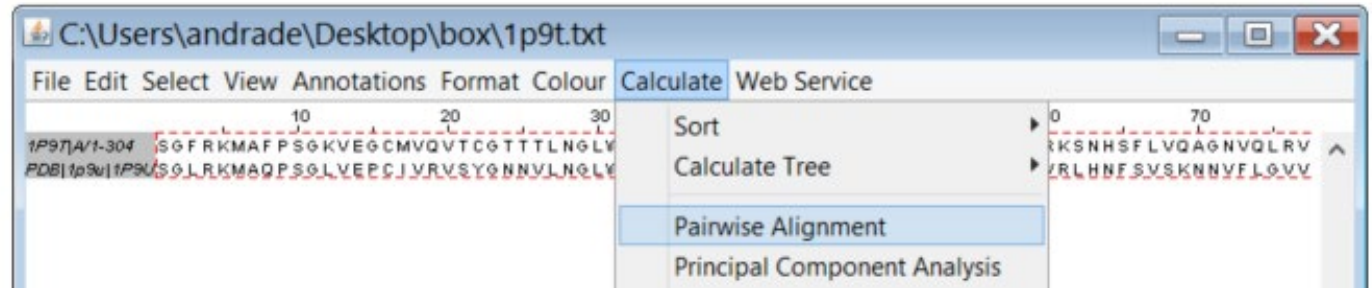
Look at the two sequences.

Do they look similar? Are they well aligned?

# SARS example

To align the sequences: select the two sequences.

Menu (of this window!) > Calculate > Pairwise alignment



What's the percentage of identity between these two sequences?

Click the button "View in alignment editor"

Can you find gaps in the alignment: How many?

Color the alignment by "percentage identity"

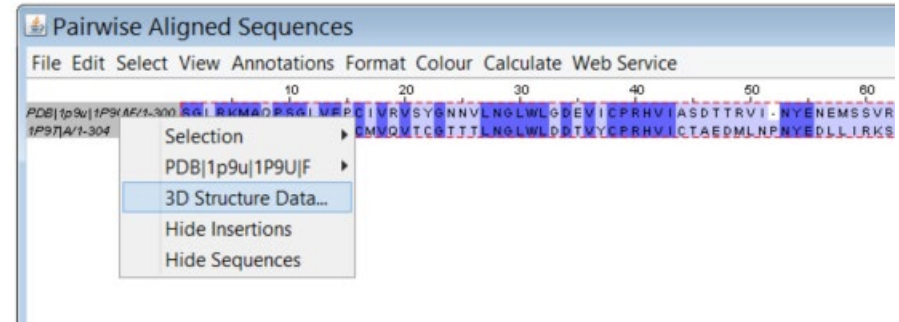
# SARS example

Now we will compare the two structures.

Select both sequences by clicking in the identifiers.

Right click on them.

Select 3D structure data...



Select both lines and click view.

What happened? :-)

# SARS example

The story continues...

Berry 2015 - virtual screening

